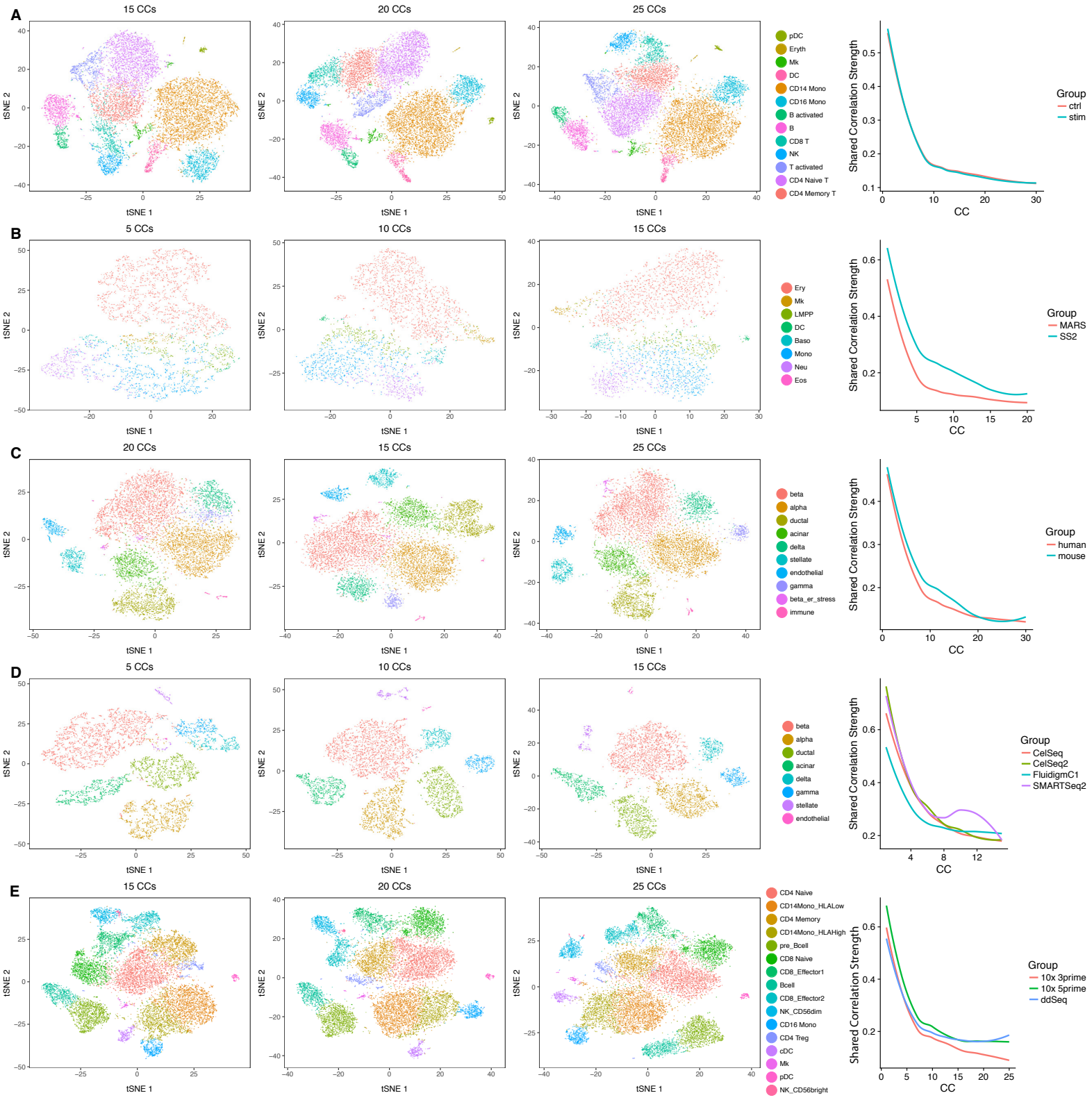
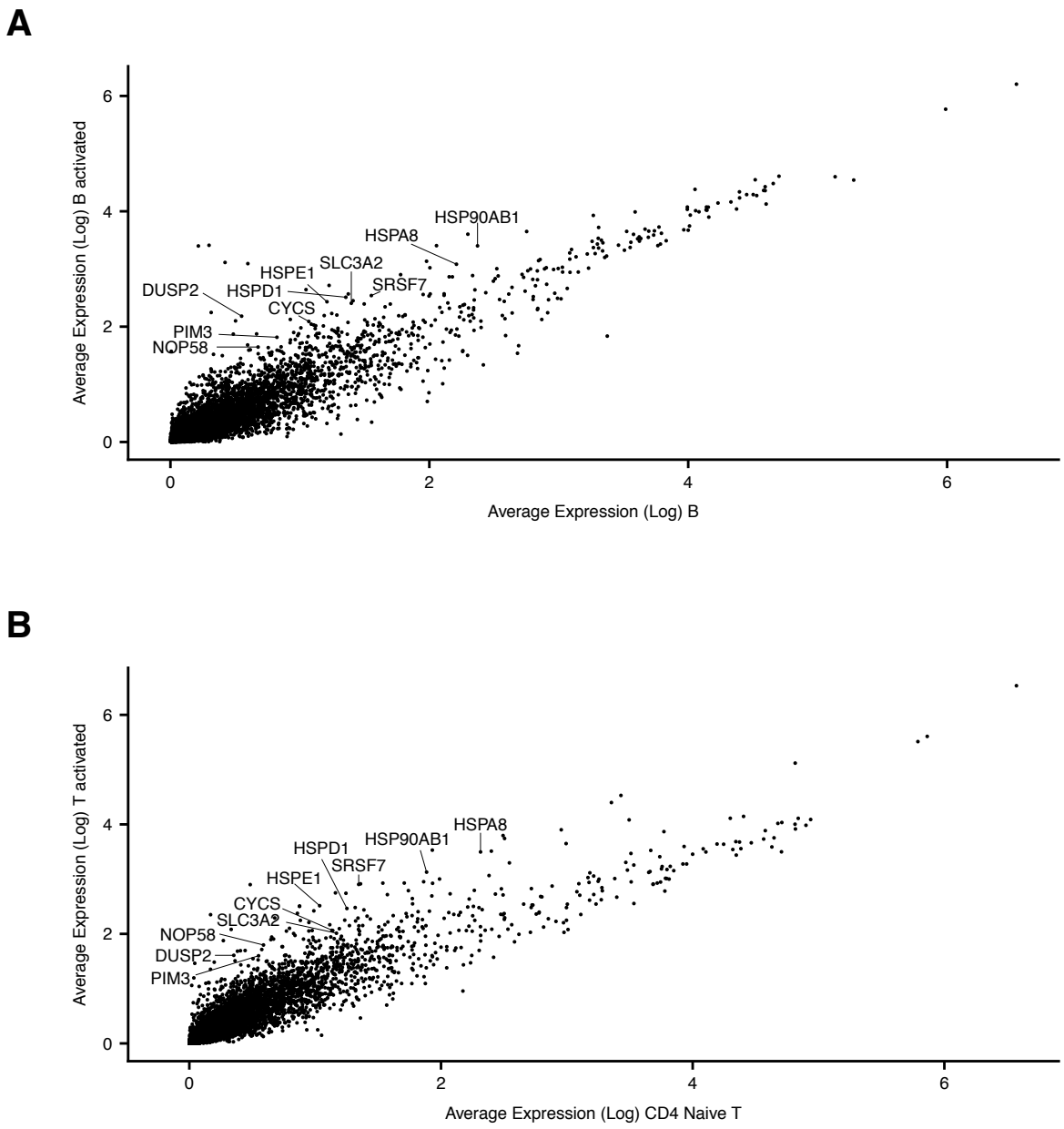


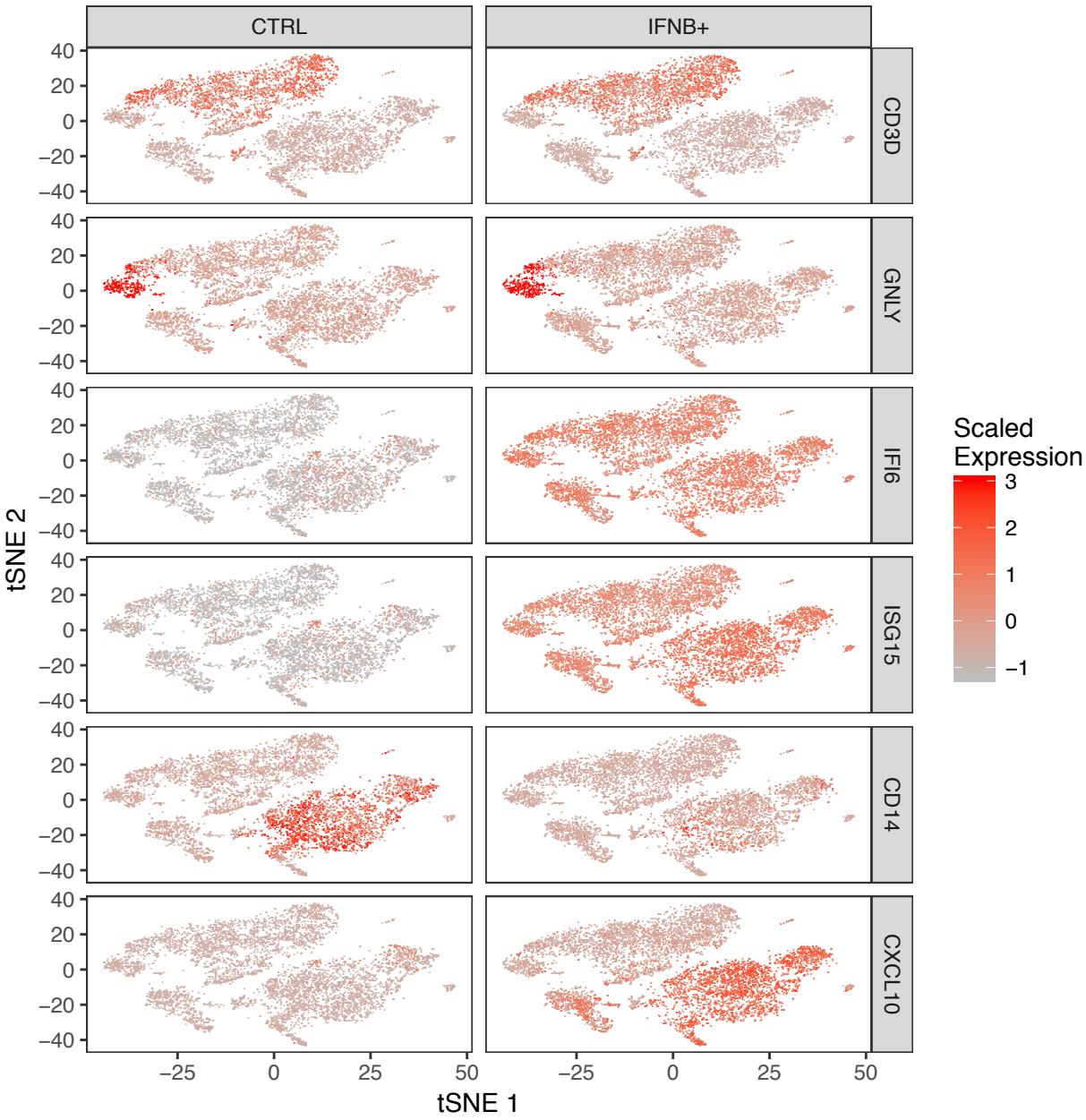
Supplementary Figure 1: CC Selection and Robustness Analysis

The Seurat integration procedure returns an aligned set of canonical correlation vectors, that can be used to construct a cell distance matrix for downstream analyses such as clustering and visualization. Choosing the dimensionality of this space (the number of CCs to include) is analogous to choosing the choosing the number of principal components to include in a standard clustering analysis. For all five examples in this manuscript, (corresponding to analyses in Figures 2, 3, 4, and Supplementary Figures 10 and 11, consisting of n = 14,039 cells, n = 3,451 cells, n = 10,306 cells, n = 6,224 cells, and n = 16,653 cells respectively),), we visualized the robustness of this parameter choice using tSNE, spanning a range of 10 CC vectors for each analysis. The plots in the second column represent the number of CCs chosen for the downstream analysis, the plots in the first column use five fewer CCs, and the plots in the third use five additional CCs. In the fourth column, we calculate the average gene biweight correlation ("Shared correlation strength"; Supplementary Methods) as a function of CC vector, to guide parameter selection. While the exact saturation point can be subjective, we observe that the global structure of our integrated dataset is robust to the exact choice of this parameter.



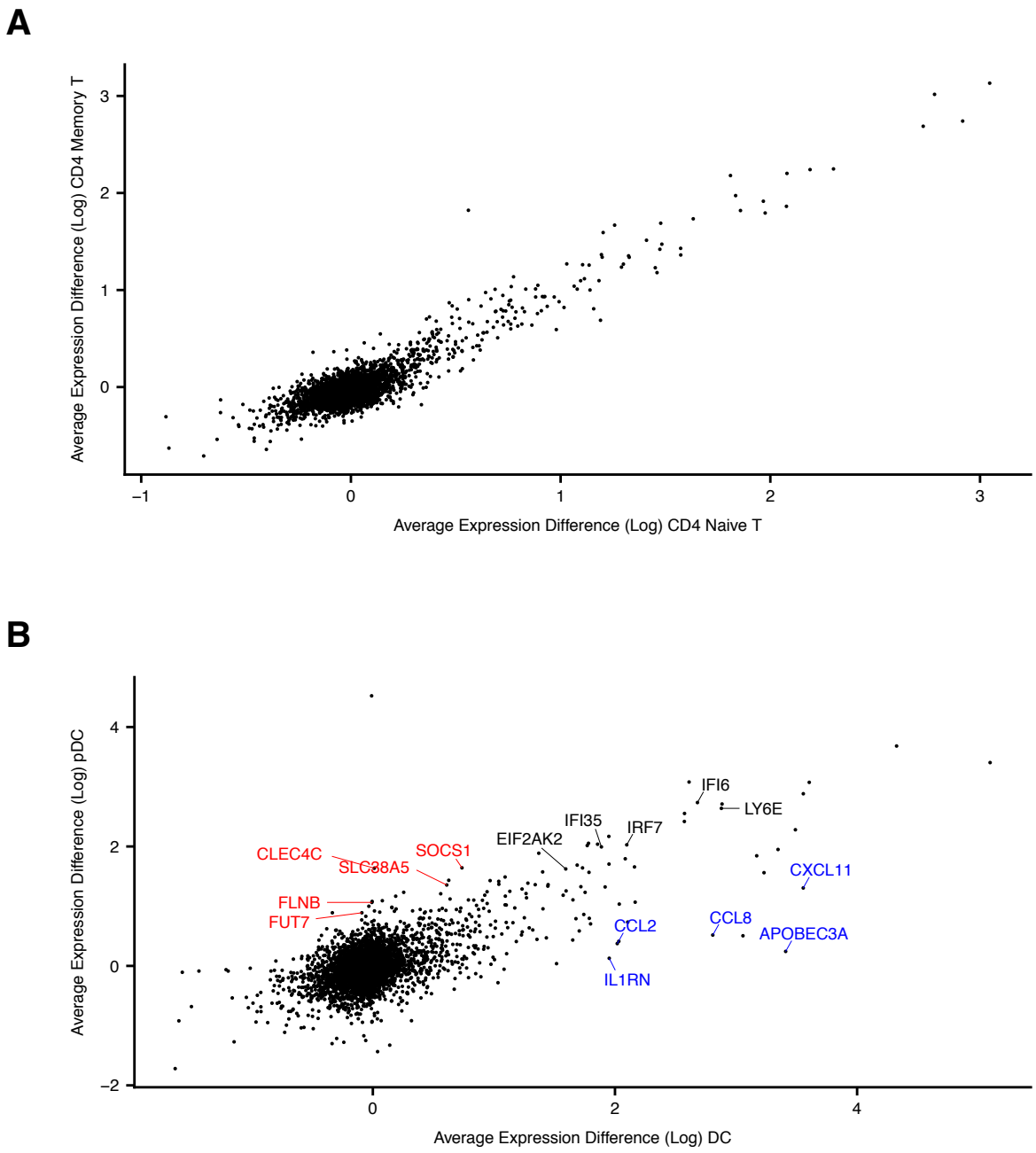
Supplementary Figure 2: Stressed cell signature in B and T cells from cell culture

Subpopulations of B and T cells are marked by the expression of heat−shock and other stress−related genes, causing them to separate in unbiased clustering. As these cells were identified both in IFN−β+ and resting cells, we hypothesize that this represents an activated stress program that is an artifact of the cell culture process. **(A−B)** Average expression per gene (n = 14,053 genes) in (A) B cells (n = 1,012 cells) versus activated B cells (n = 357 cells) and (B) CD4+ naive T cells (n = 2,608 cells) versus activated T cells (n = 787 cells). Labeled genes are upregulated in both activated B and T cells across stimulation condition.



Supplementary Figure 3: Feature plots comparing gene expression in the IFNβ+ population and the culture matched control

tSNE plots of the IFNβ stimulated (right column, n = 7,466 cells) and culture matched controls (left column, n = 6,573 cells) with cells colored by the scaled gene expression values for select genes. CD3D and GNLY are representative of cell type specific markers that don't change across interferon exposure, IFI6 and ISG15 are representative of interferon response genes that change in every cell type, and CD14 and CXCL10 are genes that also change in response to interferon but exhibit cell type specific responses. This data shows a subset of the markers displayed in Figure 2D, but with single cell resolution.



Supplementary Figure 4: Change in average gene expression across interferon stimulation condition

(A–B) Difference in average expression of genes between the population exposed to IFN–β and the culture matched control. (A) Naive (n = 2,608 cells) and memory T cells (n = 1,622 cells) displayed similar gene expression responses to interferon exposure whereas (B) DCs (n = 474 cells) and pDCs (n = 109 cells) showed cell type specific gene expression differences. Labeled genes indicated DC specific (blue), shared (black), and pDC specific (red) responses ––– genes also shown in Figure 2F.

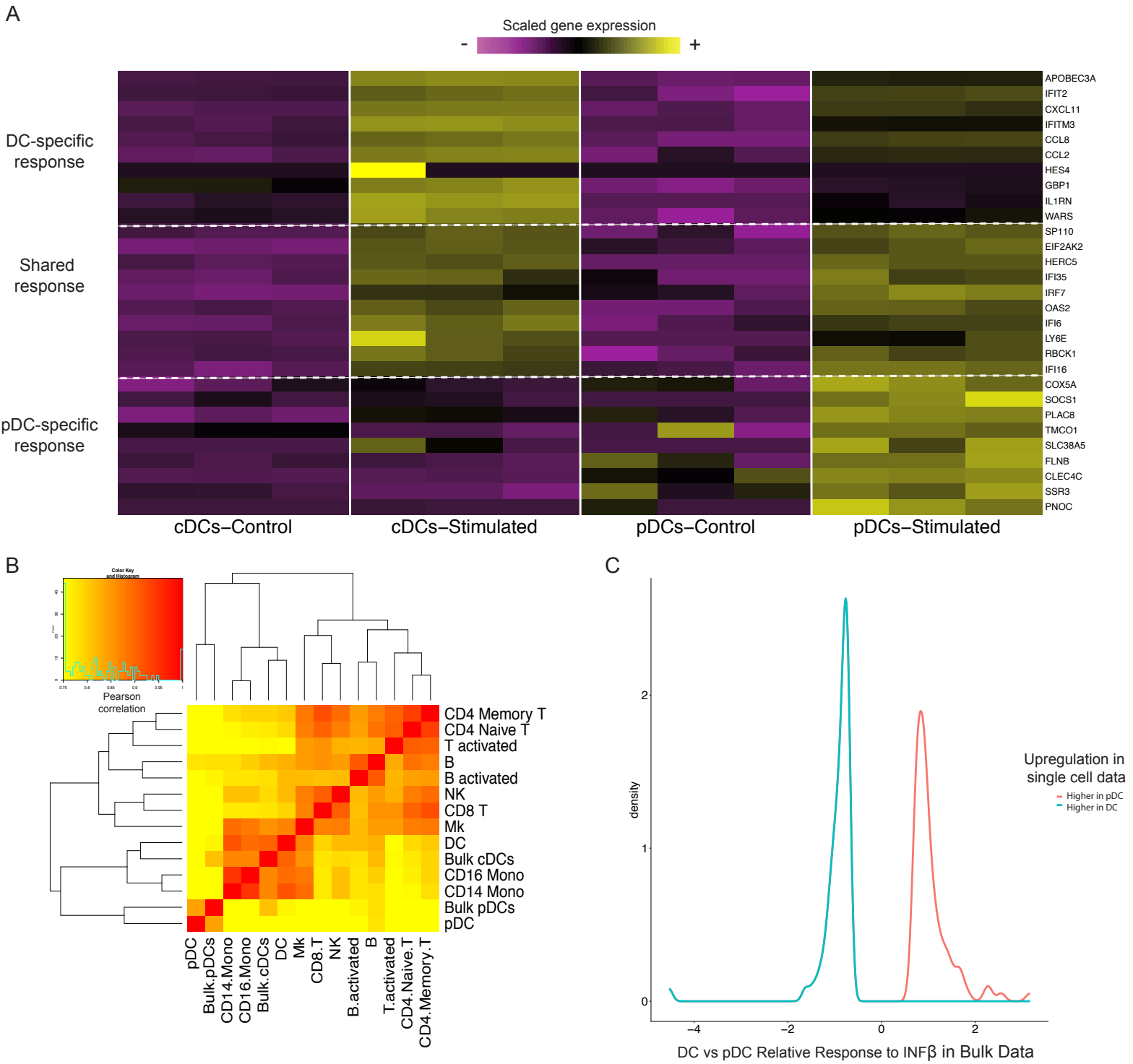
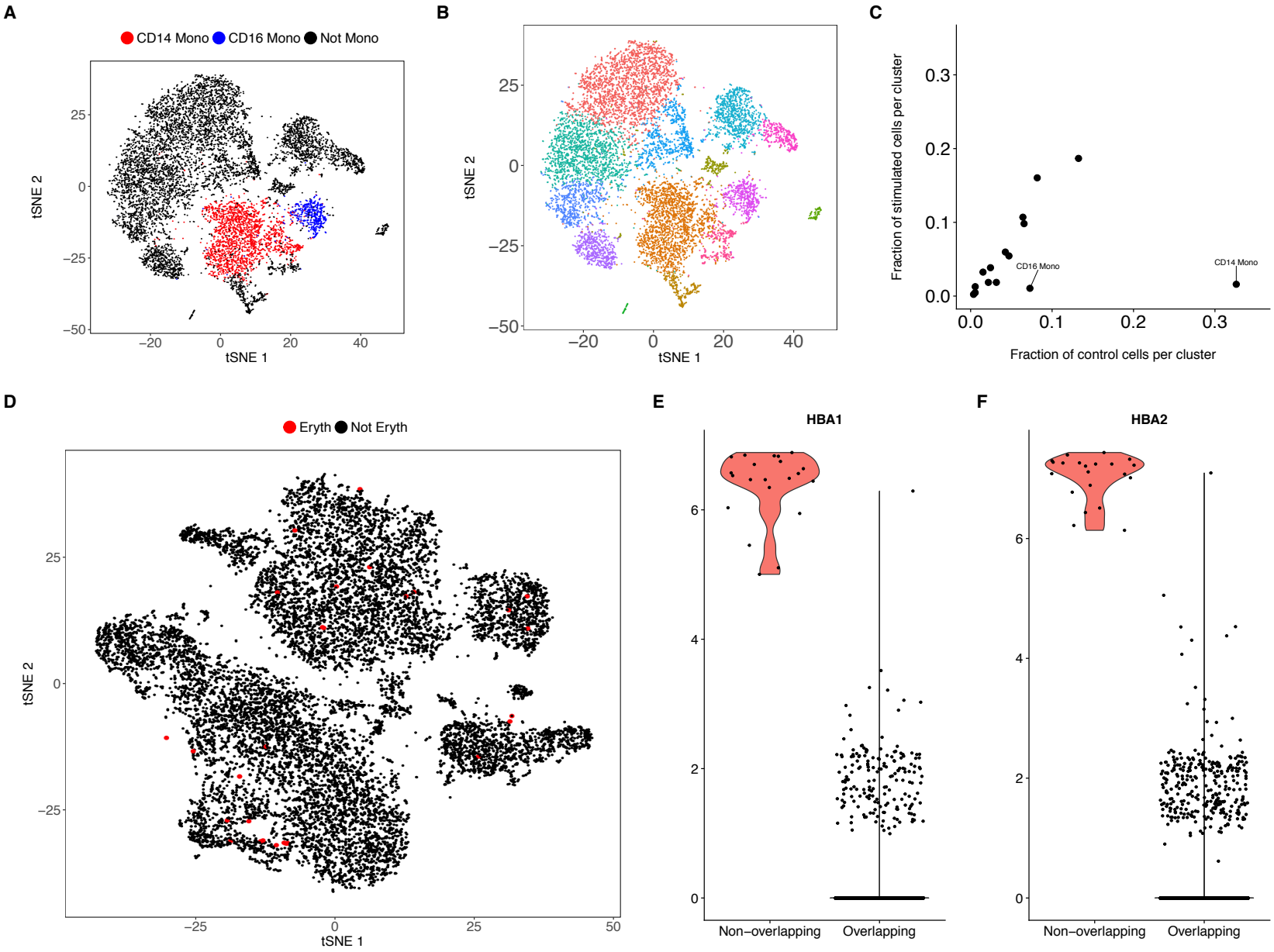


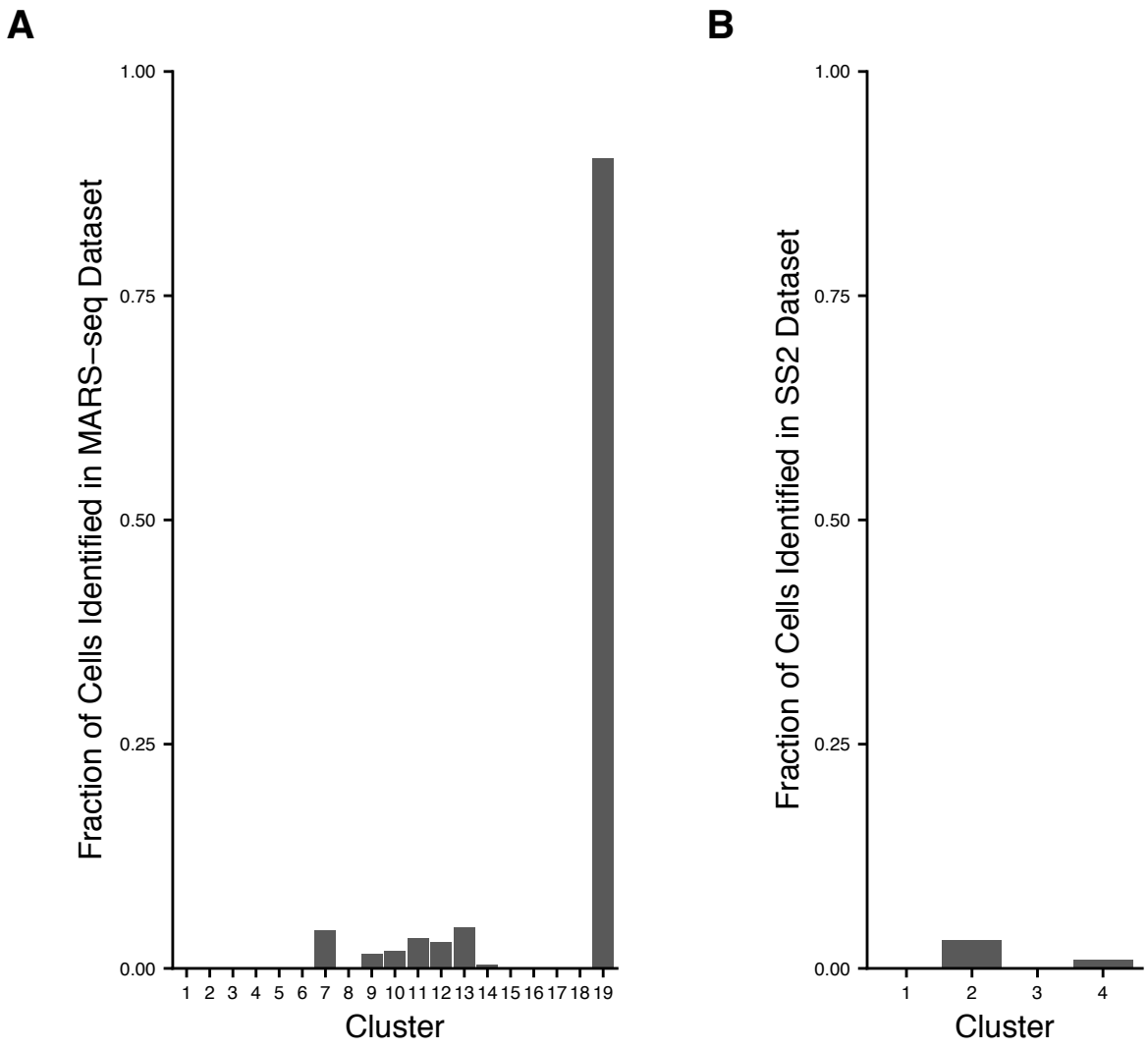
Figure S5. Validation of pDC vs. DC transcriptional responses to IFNβ

(A) Heatmap of scaled bulk RNA-seq values for selected genes in sorted pDCs and DCs after stimulation with IFNβ (Supplementary Methods). Columns correspond to technical replicates. Genes shown are the same as the single cell derived heatmap in Fig. 2F, and exhibit identical patterns. **(B)** Global analysis shows strong agreement between single cell predictions and bulk RNA-seq validation (n = 430 genes). Same as in Figure 1G, but adding in the bulk RNA-seq samples (mean of three technical replicates). pDC from the bulk samples clustered directly with the 'in silico' bulk pDC, and DC from the bulk samples cluster directly with the 'in silico' bulk DC. **(C)** For all genes (n = 153 genes) where pDC and DC exhibited different transcriptional responses to IFNβ based on our single cell predictions, we examined their expression in the bulk validation experiments, and observed identical patterns. Genes which exhibited higher IFNβ induction in DCs compared to pDCs (red curve) in the single cell data, also exhibited higher induction in the bulk data. Conversely, genes which exhibited lower induction in cDCs compared to pDCs (blue curve), also exhibited the same pattern in the bulk data. The X-axis for panel C is defined as: $log\left(\frac{DC_{stim}/DC_{ctrl}}{pDC_{stim}/pDC_{ctrl}}\right)$



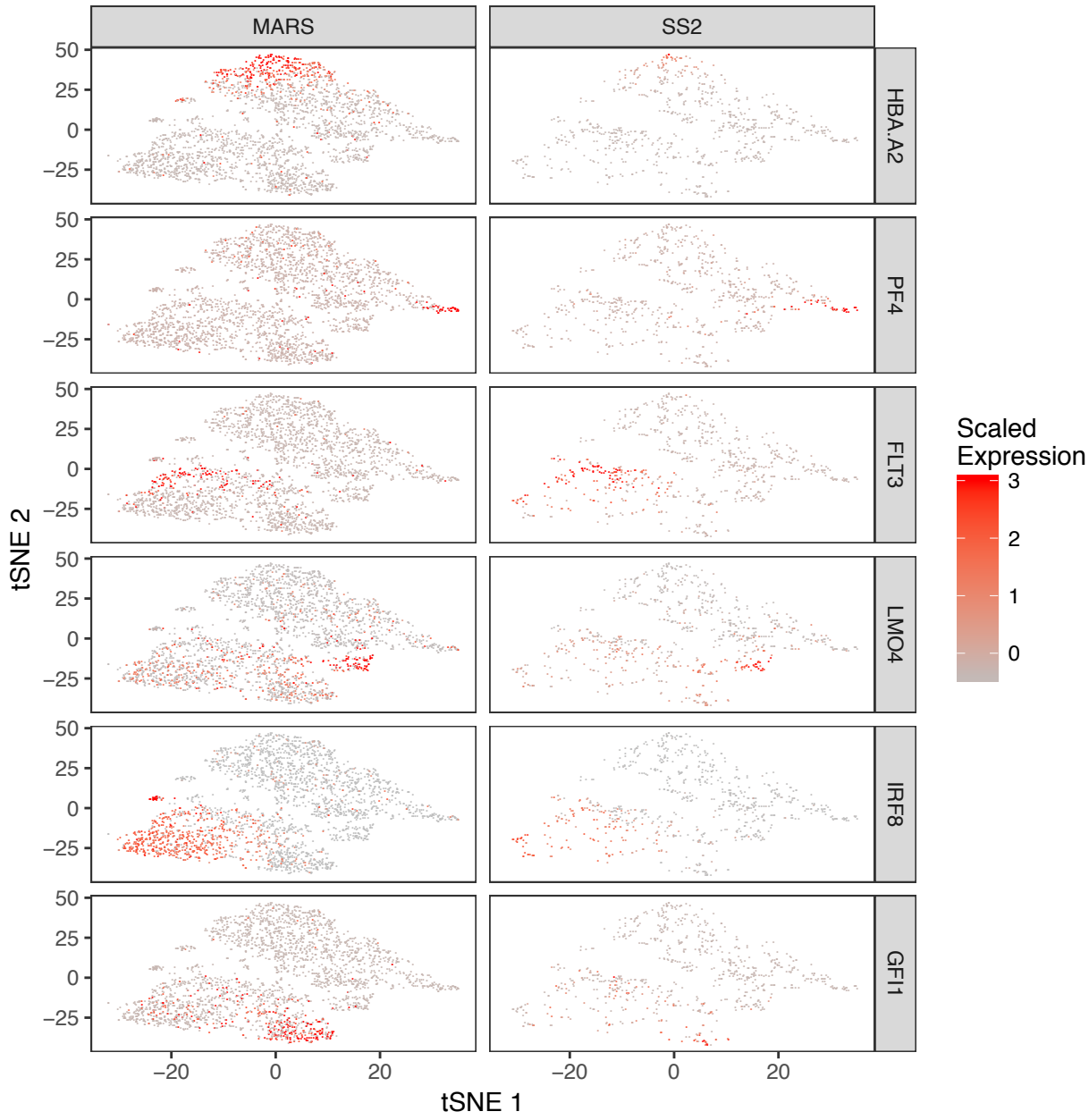
Supplementary Figure 6: Synthetic removal of specific subpopulations

To examine the robustness of our procedure to abundant non-overlapping populations, we artificially removed cells from the stimulated dataset in Kang et al. while leaving the control cells unaltered. **(A-C)** tSNE visualization (n = 11,428 cells) after the removal of all CD14+ and CD16+ monocytes from the stimulated dataset, followed by the integration procedure. **(A)** CD14+ or CD16+ monocytes from the control group continue to group together in t-SNE and **(B)** group together in graph-based clustering **(C)** In contrast to Figure 2E, the CD14+ (cluster 7) and CD16+(cluster 8) clusters now consist of control cells. **(D-F)** Similar analysis after removing erythroblasts from the stimulated group (n = 14,002 cells). **(D)** While other populations are unaffected, erythroblasts from the control cells no longer separate as a distinct group. **(E-F)** Based on the relative variance explained for each cell using PCA or CCA, we can flag individual cells as belonging to 'non-overlapping' states (n = 21 'non-overlapping' cells, n = 13,981 overlapping cells) (Supplementary Methods). Here, these cells correspond exclusively to the HBA1+ control erythroblast cells.



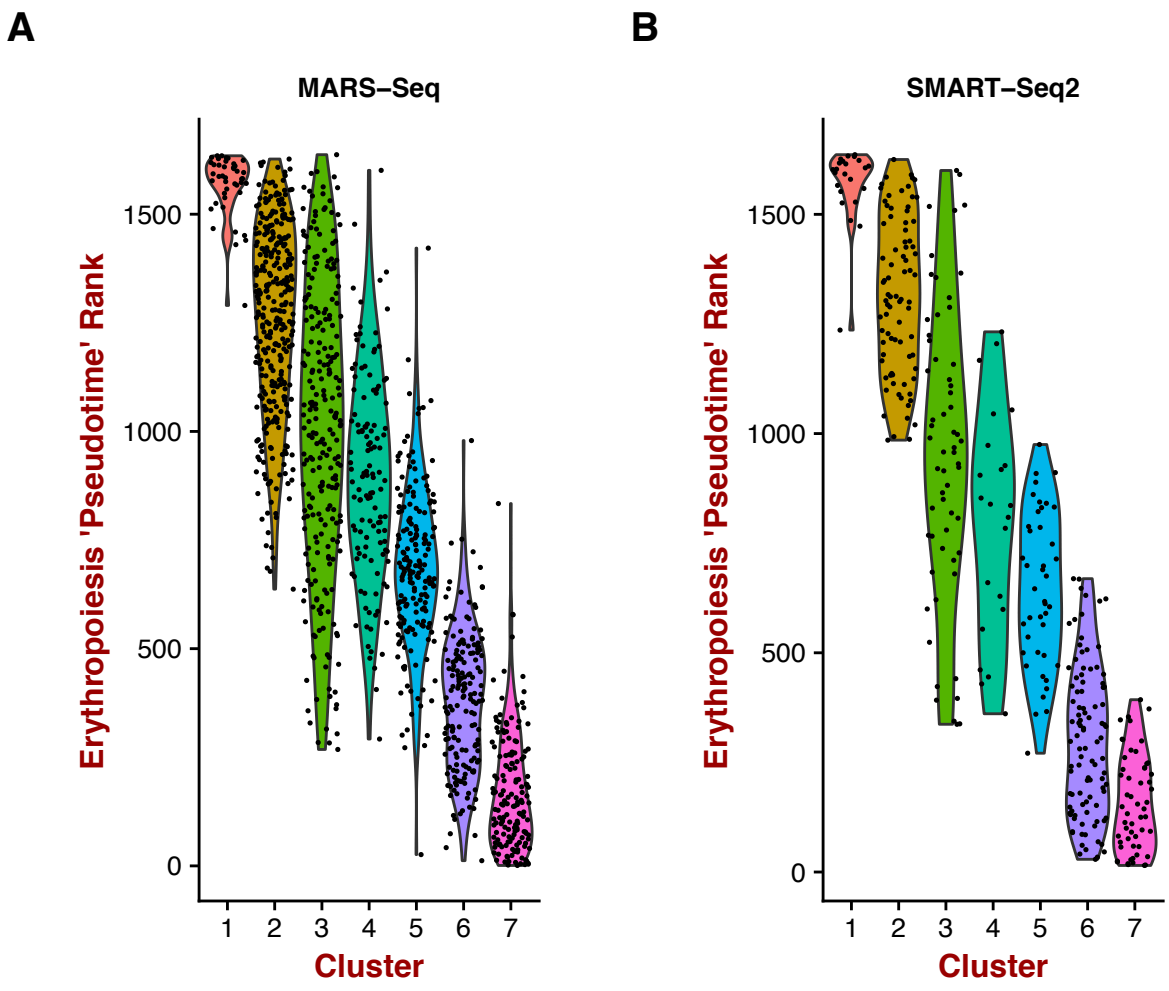
Supplementary Figure 7: Non−overlapping cell states for the hematopoietic progenitor cell datasets

(A) The fraction of cells flagged as potentially belonging to 'non−overlapping' states in each of the MARS−seq clusters is low or zero for all clusters, except the contaminating population of NK cells (cluster 19) that is unique to this dataset. **(B)** For the SS2 dataset, this is low or zero for all clusters. Cluster assignments are based on the original manuscripts.



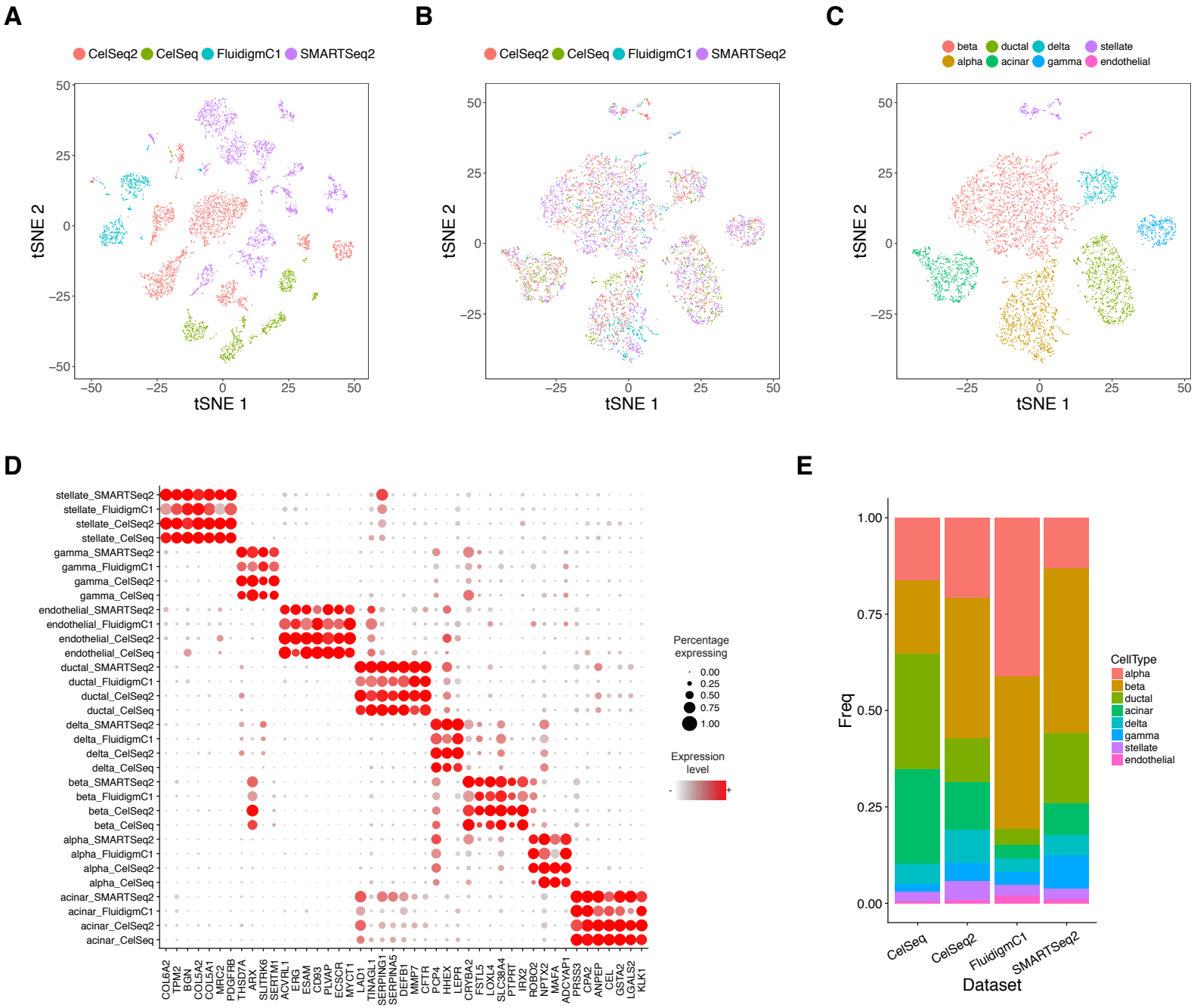
Supplementary Figure 8: Feature plots comparing gene expression in both hematopoeitic progenitor cell datasets

tSNE plots of the MARS−Seq (left column, n = 2,686 cells) and the SMART−Seq2 (right column, n = 765 cells) datasets with cells colored by the scaled gene expression values for select canonical marker genes, demonstrating the presence of the same progenitor populations in both datasets. HBA.A2 marks erythroid progenitors, PF4 marks megakaryocyte progenitors, FLT3 marks the LMPP population, LMO4 marks basophil progenitors, IRF8 marks monocyte progenitors, and GFI1 marks neutrophil progenitors. A heatmap with a greater number of lineage−specific markers is shown in Figure 3E−F.



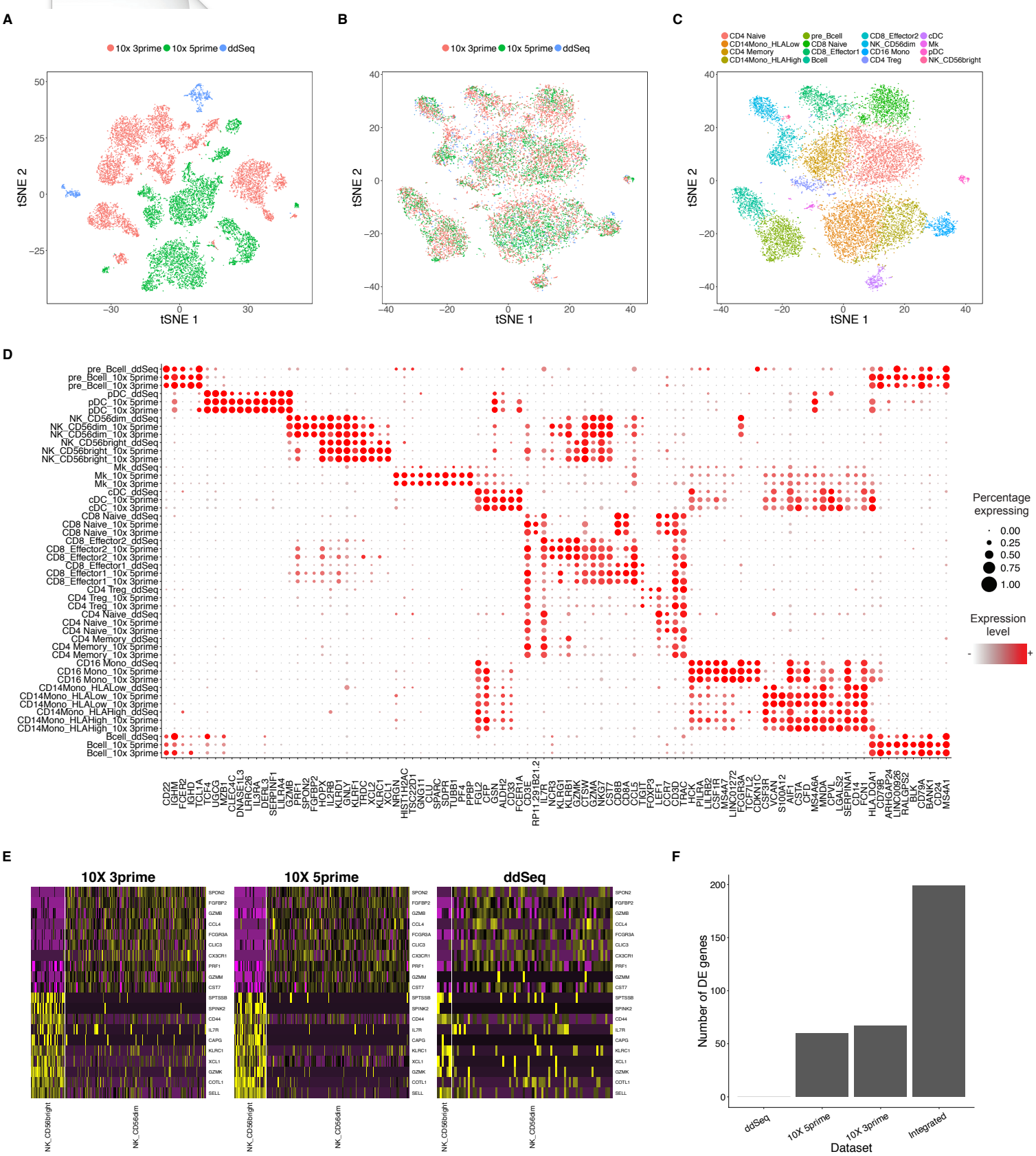
Supplementary Figure 9: Aligned trajectories of erythropoiesis

The original MARS−Seq publication defined erythropoiesis as a progression from C7 to C1, agreeing strongly with our 'pseudotemporal' orderings in both datasets. **(A−B)** For both the MARS−Seq (n = 1,255 cells) and SMART−Seq2 (n = 382 cells) datasets, cells are ranked by their projection onto a principal curve fit through the joint diffusion map modeling erythropoiesis (Figure 3G). Cells are grouped by either the original cluster assignments for the MARS−Seq data (A) or by the MARS−Seq cluster they map to (B).



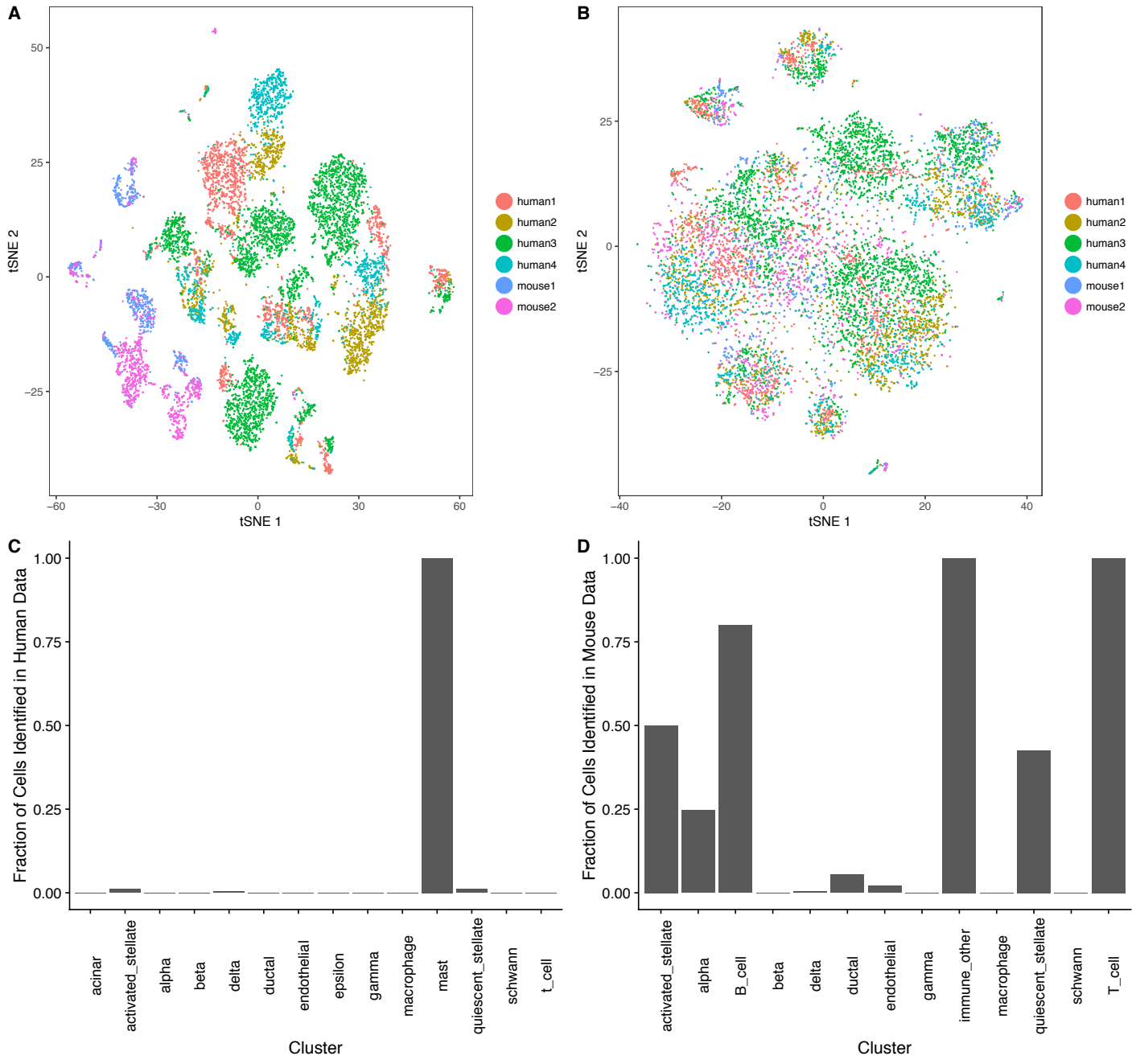
Supplementary Figure 10: Integration of scRNA-seq datasets from complex tissues across multiple technologies.

(A-C) tSNE plots of human pancreatic islet cells (n = 6,244 cells) profiled with four scRNA-seq technologies pre **(A)** and post **(B)** alignment. After alignment, cells across technologies conditions group together based on shared cell type, allowing for a single joint clustering **(C)** to detect 8 endocrine, exocrine, endothelial, and stellate populations. **(D)** Integrated analysis reveals clear markers of cell types that were conserved across all four technologies **(E)** The composition of cell types varies widely between donors, but our procedure is robust to this.



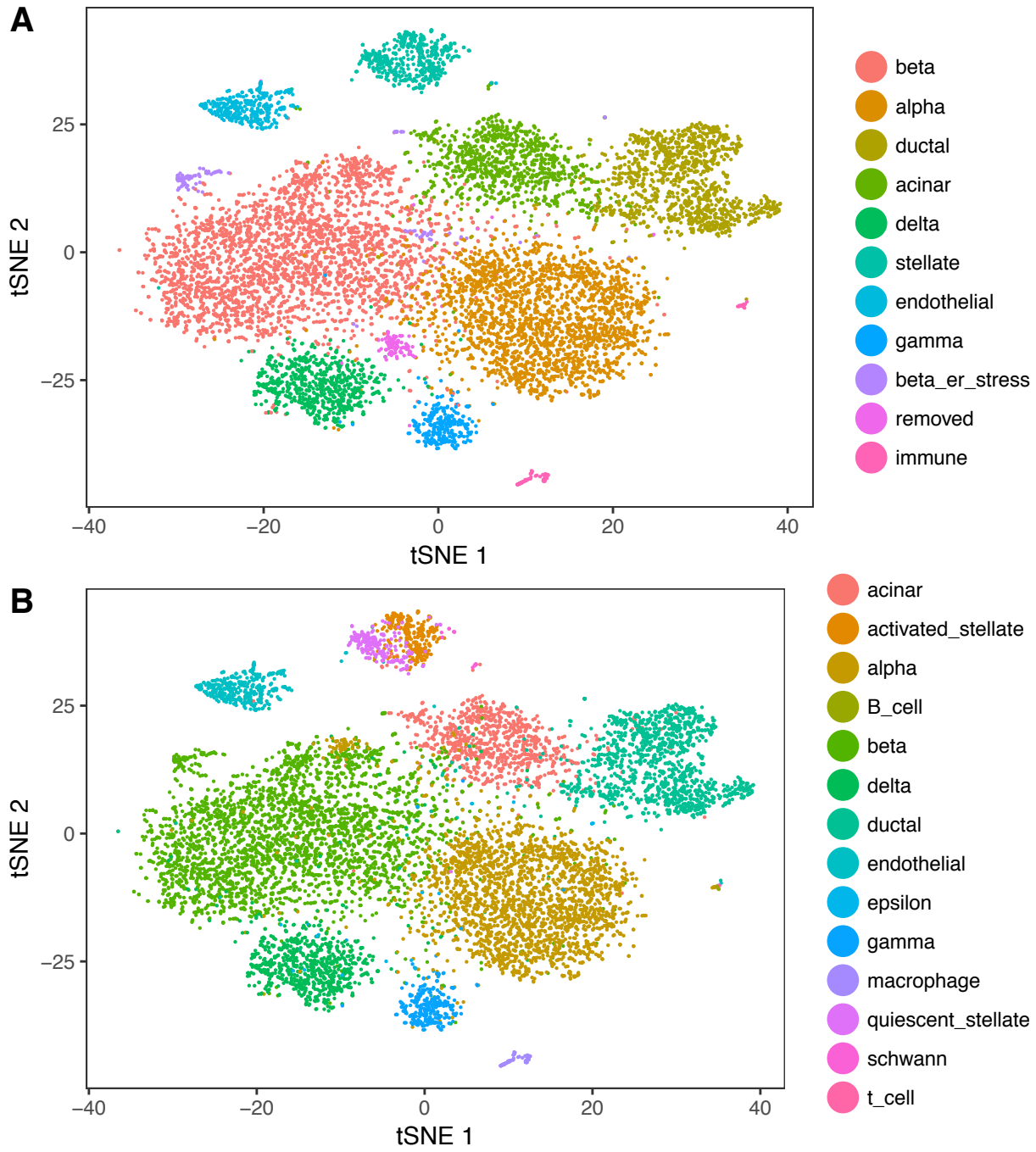
Supplementary Figure 11: Integration of PBMC scRNA-seq datasets across multiple technologies.

(A-C) tSNE plots of human PBMCs (n = 16,653 cells) profiled with three scRNA-seq technologies pre (A) and post (B) alignment. Joint clustering after integration reveals 16 immune populations (C) including rare populations of CD56^{bright} NK cells. **(D)** Integrated analysis reveals clear markers of cell types that were conserved across all three technologies**. (E)** Single cell heatmaps showing a subset of differentially expressed genes between CD56^{bright} and CD56^{dim} cells that are conserved across technologies (at most 150 cells are shown per cluster for visualization). **(F)** Notably, a 'meta-analysis' of differential expression that pooled cells across all three technologies was able to return almost three times the number of DE genes based on a p-value threshold of 10^{-5}, compared to individual analysis within any technology. No genes were identified as DE in the ddSeq data alone, as only 6 cells were identified as CD56^{bright}.



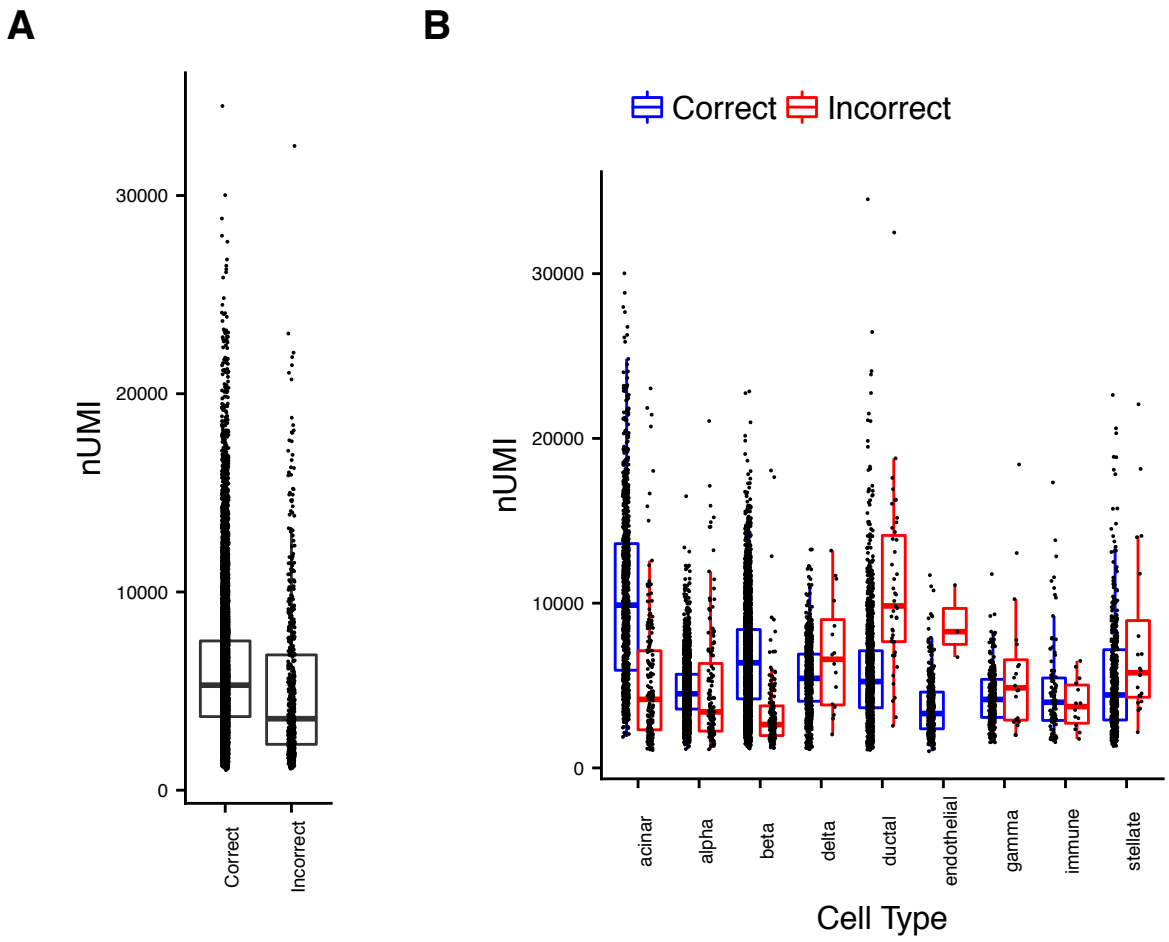
Supplementary Figure 12: Donor effects and non–overlapping cell state identification in the pancreatic islet cell datasets

Prior to alignment, pancreatic islet cells (n = 10,191 cells) group by cell type, species, and also individual donor **(A)**. After alignment (**B**) both species and donor effects are largely mitigated. Though this appears to result in a loss of structure in the dataset post–alignment, this is due to the fact that individual–specific signals do not replicate across species, and are therefore not captured by the alignment procedure. Rare immune subpopulations, including mast cells in the human dataset **(C)** and B and T cells in the murine dataset **(D)** are identified as potentially originating from 'non–overlapping' states are indeed present in only one of the two datasets. Cells are grouped here by the original assignments in Baron et al 2016.



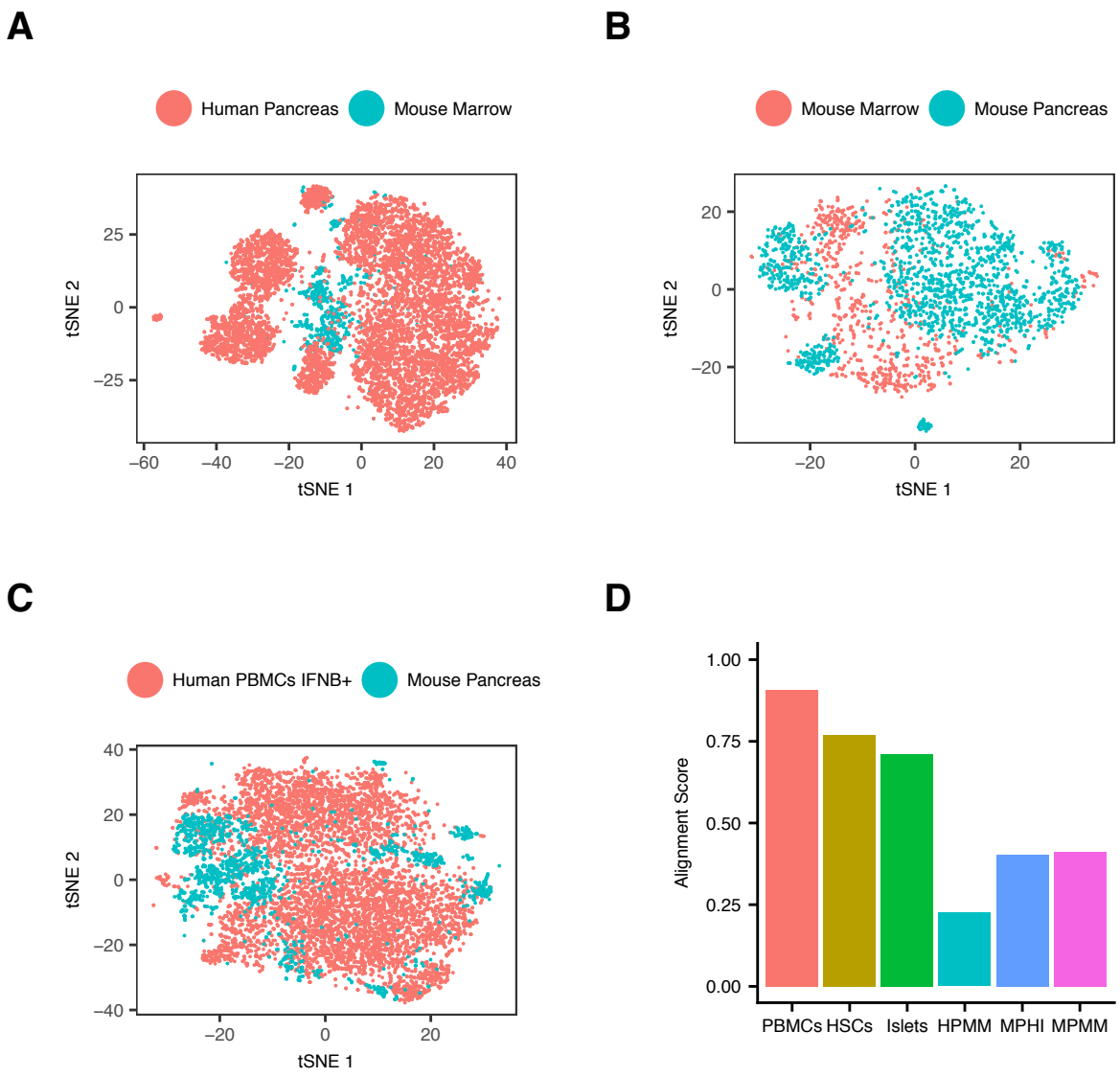
Supplementary Figure 13: Pancreas islet cell type assignments

(A) tSNE plot of all pancreatic islet cells (n = 10,306 cells) post–alignment, colored by cell type as determined by unsupervised clustering. This figure is identical to 4C, but includes an 11th cluster of 115 cells that were defined by low–complexity and gene counts (median 1,088 genes) in both species, which we removed from downstream analyses. **(B)** tSNE plot, with the same positioning (post–alignment), but displaying the original cell type identities from Baron et al. 2016 (n = 10,191 cells).



Supplementary Figure 14: Number of UMIs detected per cell for correctly and incorrectly classified islet cells

(A–B) Our unbiased clustering, performed jointly across species, agrees overwhelmingly with the analysis of original datasets from Baron et al. (9594 concordant classifications, 597 discordant classifications). The majority of discordant calls (394/597) represented acinar, alpha, and beta cells, where discordant cells exhibited significantly lower UMI counts than concordant cells (two–sided Wilcoxon rank sum test p< 2.2e–16, p=3.419e–5, and p<2.2e–16 respectively). The interquartile range (IQR) is captured by the hinges on each boxplot and whiskers extend to at most 1.5 * IQR of each hinge. Summary of data distributions and statistical details can be found in Supplementary Table 4.



Supplementary Figure 15: Negative controls – integrated analysis of biologically dissimilar tissues

As a negative control for the alignment procedure, we attempted to align three dataset pairs that should have minimal biological similarity. **(A–C)** The tSNE plots with cells colored by dataset after alignment for each of the three negative controls demonstrate that cells separate almost entirely by the dataset they belong to. In (A), we tried to align human pancreatic islet cells (n = 8,456 cells) and mouse hematopoetic progenitor cells (n = 765 cells) ––– HPMM. In (B) we tried to align mouse hematopoetic progenitor cells (n = 760 cells) and mouse pancreatic islet cells (n = 1,733 cells) ––– MPMM. Finally, in (C) we tried to align stimulated human PBMCs (n = 7,270 cells) with mouse pancreatic islet cells (n = 1,770 cells) ––– MPHI. **(D)** We also computed alignment scores for each of these three negative controls, which were substantially lower than for the datasets with shared cell types.